/* PROBLEM SET 6 ANSWERS – SPECIFICATION ANALYSIS */

. u ps4data    /* read in data */
. g lhpay=log(hourpay)  /* create log of hourly pay */
. g age2=age^2   /* create age squared */

/* good idea always to inspect your data by looking at mean values and variance of variables you will use */

. su lhpay age age2 parttime postgrad grad highint none tenure

Variable |     Obs     Mean   Std. Dev.       Min       Max
---------+-----------------------------------------------------
lhpay | 12098  1.969559   .5713568  -3.506558   5.602119
age | 12098  39.0996    11.19333         16         64
age2 | 12098 1654.059    890.2529        256       4096
parttime | 12098   .2301207  .4209273          0          1
postgrad | 12098   .0517441  .2215191          0          1
grad | 12098  .1270458    .333038          0          1
highint | 12098  .1849893   .3883049          0          1
none | 12098  .1129112   .3164973          0          1
tenure | 12098  91.98454   96.89718          0        576

Can see average age is 39, 23% work part-time, 12.7% are graduates, 11% have no qualifications

. reg lhpay age age2 parttime

Source |       SS      df       MS                    Number of obs =   12098
---------+------------------------------                    F(  3, 12094) =  767.96
Model |  631.910209     3  210.636736                    Prob > F      =  0.0000
Residual |   3317.1389 12094  .274279717                   R-squared     =  0.1600
---------+------------------------------                   Adj R-squared =  0.1598
Total |  3949.04911 12097  .326448633                   Root MSE      =  .52372

------------------------------------------------------------------------------
 lhpay |      Coef.    Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------
age |  .0900186    .0028852    31.200   0.000        .0843632    .0956741
age2 |  -.0010387   .0000363    -28.644   0.000       -.0011098   -.0009676
parttime |  -.4032982   .0113554    -35.516   0.000      -.4255566   -.3810398
_cons |   .2607226    .0543607     4.796   0.000        .154167    .3672783
------------------------------------------------------------------------------

Regression is log-lin so estimated coefficients give percentage increase in hourly pay for unit change in variable

When age is entered as a quadratic (age and age squared together) then need to differentiate

$$\ln W = a + b \times \text{age} + c \times \text{age}^2 + d \times \text{parttime}$$

To get total effect of age

So

$$\frac{d\ln W}{d\text{age}} = b + 2c \times \text{age}$$

Says that partial effect of age is not constant but varies with age

So if age=16

$$\frac{d\ln W}{d\text{age}} = .09 + 2(-.001\times16) = .058$$

ie hourly pay rises by 5.8% if age 1 extra year
If age = 30  \[ \frac{d\ln w}{d\text{Age}} = 0.09 + 2(-0.001 \times 30) = 0.030 \]

ie hourly pay rises by 3.0% if age 1 extra year

If age = 30  \[ \frac{d\ln w}{d\text{Age}} = 0.09 + 2(-0.001 \times 60) = -0.030 \]

ie hourly pay falls by 3.0% if age 1 extra year

Age is maximised when \( \frac{d\ln w}{d\text{Age}} = 0 \) (1st order condition for a maximum)

So from (1)  \[ \frac{d\ln w}{d\text{Age}} = b + 2c \times \text{age} = 0 \]

\[ 0.09 + 2 \times (-0.001 \times \text{age}) = 0 \]

Solving for age implies  \[ \text{age} = \frac{0.09}{0.002} = 45 \]

ie suggests earnings reach a peak at age 45 years.

For dummy variables then coefficient is an intercept shift (changes the constant in the regression line if and only if the individual has that characteristic)

So from regression output predicted earnings are

\[
\begin{align*}
\text{if parttime} = 1 & : 0.26 + 0.09\times \text{Age} - 0.001 \times \text{Age}^2 - 0.403 \\
\text{if parttime} = 0 & : 0.26 + 0.09\times \text{Age} - 0.001 \times \text{Age}^2
\end{align*}
\]

Since log-lin model suggests average hourly pay of part-time workers is 40.3% lower.

Now introduce interactive dummy variable (product of a dummy variable and a continuous variable)

\[ \text{g ptage} = \text{parttime} \times \text{age} \]

\[ \text/* this then implies a different age effect for those who have this characteristic \]

\[ \text{reg lhpay age age2 parttime ptage} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 12098</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>637.87443</td>
<td>4</td>
<td>159.468607</td>
<td>F( 4, 12093) = 582.41</td>
</tr>
<tr>
<td>Model</td>
<td>3311.17468</td>
<td>12093</td>
<td>.273809202</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>3949.04911</td>
<td>12097</td>
<td>.326448633</td>
<td>R-squared = 0.1615</td>
</tr>
<tr>
<td></td>
<td>12097</td>
<td>.326448633</td>
<td>Root MSE = 0.52327</td>
<td></td>
</tr>
</tbody>
</table>

| lhpay | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|---|-----|------------------------|
| age   | 0.0901803 | 0.0028829 | 31.281 | 0.000 | 0.0845293 | 0.0958313 |
| age2  | -0.0010274 | 0.0000363 | -28.295 | 0.000 | -0.0010986 | -0.0009563 |
| parttime | -0.2077354 | 0.0434107 | -4.785 | 0.000 | -0.2928274 | -0.1226435 |
| ptage | -0.0048479 | 0.0010387 | -4.667 | 0.000 | -0.006884 | -0.0028118 |
| _cons | 0.2362985 | 0.0545656 | 4.331 | 0.000 | 0.1293412 | 0.3432557 |

So now initial part-time/full-time intercept difference is -20.8% but this rises by 0.5% a year as individuals get older, (slope effects for full and part-timers drift apart)

Now include education dummy variables (all intercept dummies so estimated coefficients give average difference between the groups relative to the missing (default category)
### Regression Output

#### Model Summary

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 12098</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1301.14126</td>
<td>8</td>
<td>162.642657</td>
<td>F( 8, 12089) = 742.54</td>
</tr>
<tr>
<td>Residual</td>
<td>2647.90785</td>
<td>12089</td>
<td>.219034482</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3949.04911</td>
<td>12097</td>
<td>.326448633</td>
<td>R-squared = 0.3295</td>
</tr>
</tbody>
</table>

#### Coefficients

| lhpay | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|-------|------|----------------------|
| age   | 0.0787139 | 0.0025881 | 30.414 | 0.000 | 0.0736408 - 0.0837871 |
| age2  | -0.0008466 | 0.0000327 | -25.928 | 0.000 | -0.0009106 - -0.0007826 |
| parttime | -0.0996061 | 0.0389108 | -2.560 | 0.010 | -0.1758775 - -0.0233348 |
| ptage | -0.0056442 | 0.0009307 | -6.065 | 0.000 | -0.0074685 - -0.00382 |
| postgrad | 0.8144316 | 0.0229259 | 35.525 | 0.000 | 0.7694931 - 0.85937 |
| grad | 0.7470994 | 0.0179397 | 41.645 | 0.000 | 0.7119348 - 0.782264 |
| highint | 0.4769245 | 0.0212329 | 29.062 | 0.000 | 0.4447571 - 0.5090918 |
| low | 0.2152263 | 0.0054271 | 15.067 | 0.000 | 0.1872263 - 0.2432262 |
| _cons | 0.030198 | 0.0504271 | 0.599 | 0.549 | -0.0686471 - 0.1290432 |

#### Notes

- Constant is average wage of default group (no quals) at a notional age of zero.
- Estimated postgrad coefficient says, on average postgrads earn 81.4% more than those with no quals (and 6.7% points more than graduates - 0.814 - 0.747 = 0.067).

#### Dummy Variable Trap

- Dummy variable trap says **never include** as many dummy variables as there are categories since this generates perfect collinearity between variables. (see lecture notes)
- When this is done STATA automatically drops one of the dummies.

### Regression Output with None

#### Model Summary

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 12098</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1301.14126</td>
<td>8</td>
<td>162.642657</td>
<td>F( 8, 12089) = 742.54</td>
</tr>
<tr>
<td>Residual</td>
<td>2647.90785</td>
<td>12089</td>
<td>.219034482</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3949.04911</td>
<td>12097</td>
<td>.326448633</td>
<td>R-squared = 0.3295</td>
</tr>
</tbody>
</table>

#### Coefficients

| lhpay | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|-------|-----------|-------|------|----------------------|
| age   | 0.0787139 | 0.0025881 | 30.414 | 0.000 | 0.0736408 - 0.0837871 |
| age2  | -0.0008466 | 0.0000327 | -25.928 | 0.000 | -0.0009106 - -0.0007826 |
| parttime | -0.0996061 | 0.0389108 | -2.560 | 0.010 | -0.1758775 - -0.0233348 |
| ptage | -0.0056442 | 0.0009307 | -6.065 | 0.000 | -0.0074685 - -0.00382 |
| postgrad | 0.8144316 | 0.0229259 | 35.525 | 0.000 | 0.7694931 - 0.85937 |
| grad | 0.7470994 | 0.0179397 | 41.645 | 0.000 | 0.7119348 - 0.782264 |
| highint | 0.4769245 | 0.0212329 | 29.062 | 0.000 | 0.4447571 - 0.5090918 |
| low | 0.2152263 | 0.0054271 | 15.067 | 0.000 | 0.1872263 - 0.2432262 |
| _cons | 0.030198 | 0.0504271 | 0.599 | 0.549 | -0.0686471 - 0.1290432 |

### Notes

- Changing the default category changes the reference point. Now the missing group is postgraduates, so this is the comparator group. Note the value of the constant changes because of this.
So now those with no qualifications earn 81.4% less than the reference group, (same effect but in reverse). Note estimates on age and parttime are unchanged.

Omitted variable Bias tests

. g male=sex==1    /* male dummy variable */

. reg lhpay male

 says men earn 25.5% more than women on average
Because men have on average longer job tenure, some of effect of long job tenure on earnings is picked up by male coefficient in absence of tenure variable. When tenure included, size of male coefficient falls. Positive correlation between male and tenure and positive effect of tenure on pay means that male coefficient omitting tenure is biased upward.

```
. reg lhpay male tenure age age2 numkids single
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>621.536418</td>
<td>6</td>
<td>103.589403</td>
</tr>
<tr>
<td>Residual</td>
<td>3327.51269</td>
<td>12091</td>
<td>.275205748</td>
</tr>
<tr>
<td>Total</td>
<td>3949.04911</td>
<td>12097</td>
<td>.326448633</td>
</tr>
</tbody>
</table>

| lhpay | Coef.   | Std. Err. | t       | P>|t|    | [95% Conf. Interval] |
|-------|---------|-----------|---------|--------|---------------------|
| male  | .2472006 | .0096507  | 25.615  | 0.000  | .2282837 - .2661175 |
| tenure| .00109   | .000055   | 19.806  | 0.000  | .0009821 - .0011979 |
| age   | .0890994 | .0032743  | 27.212  | 0.000  | .0826814 - .0955175 |
| age2  | -.001061 | .00004    | -27.655 | 0.000  | -.0011845 - -.0010277 |
| numkids| -.0303295| .0052991  | -5.724  | 0.000  | -.0407166 - -.0199425 |
| single| -.0079765| .0145325  | -0.549  | 0.583  | -.0364625 - .0205095 |
| _cons | .1175128 | .0653783  | 1.797   | 0.072  | -.0106392 - .2456647 |

Note single variable is insignificant. Single may be an irrelevant variable in this context.

```
. reg lhpay male tenure age age2 numkids
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>621.453509</td>
<td>5</td>
<td>124.290702</td>
</tr>
<tr>
<td>Residual</td>
<td>3327.5956</td>
<td>12092</td>
<td>.275189845</td>
</tr>
<tr>
<td>Total</td>
<td>3949.04911</td>
<td>12097</td>
<td>.326448633</td>
</tr>
</tbody>
</table>

| lhpay | Coef.   | Std. Err. | t       | P>|t|    | [95% Conf. Interval] |
|-------|---------|-----------|---------|--------|---------------------|
| male  | .2468877 | .0096336  | 25.628  | 0.000  | .2280043 - .265771 |
| tenure| .0010896 | .000055   | 19.801  | 0.000  | .0009817 - .0011975 |
| age   | .089703  | .003084   | 29.087  | 0.000  | .0836580 - .0957481 |
| age2  | -.0011113| .0000388  | -28.609 | 0.000  | -.0011875 - -.0010352 |
| numkids| -.0294023| .0050225  | -5.854  | 0.000  | -.0392471 - -.0195575 |
| _cons | .0998263 | .0568841  | 1.755   | 0.079  | -.0116757 - .2113283 |

Dropping single variable makes little difference to other estimates and both adjusted and unadjusted r squared. (Insignificant variables with t value less than one will not increase the adjusted r-squared). Standard errors of all remaining variables are smaller, (see lecture notes for reasons why)

If there
Ramsey Reset test
Test of specification error and/or omitted variable bias.

```
. predict yhat
```

```
. g yhat2=yhat^2 /* square of yhat */
```
. g yhat3=yhatˆ3
   /* cube of yhat */

. reg lhpay male tenure age age2 numkids yhat2 yhat3

Source | SS      df       MS                  Number of obs =  12098
---------+------------------------------               F(  7, 12090) =  330.08
Model    | 633.622715     7  90.5175307               Prob > F      =  0.0000
Residual | 3315.4264 12090  .274228817               R-squared     =  0.1604
---------+------------------------------               Adj R-squared =  0.1600
Total    | 3949.04911 12097  .326448633               Root MSE      =  .52367

        |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   male   |  -1.954901    .546238     -3.579   0.000      -3.025615   -.8841872
 tenure  |  -0.0084616   .002401     -3.524   0.000      -0.0131679  -.0037552
   age    |  -0.7160179   .1977751     -3.620   0.000      -1.103689   -.328347   
  age2    |   0.008627    .0024508      3.616   0.000       .0040588   .0136666
  numkids |   0.2329955   .0647653      3.598   0.000       .1060451   .3599459
   yhat2  |   5.103833    1.157005      4.411   0.000       2.835918    7.371748
  yhat3   |  -0.9511138   .2004987     -4.744   0.000      -1.344123   -.5581042
 _cons   |   4.371823    1.166719      3.747   0.000       2.084868    6.658779

12098 cases in.
12090 samples included, adjustments for 8DF missing-independence groups.

F test of null hypothesis that coefficients on Yhat2 and Yhat3 are zero
(variables have no explanatory power and so H0: no omitted variables)

F = RSSrestrict - RSSunrestrict /J ~ F(J, N-Kunrestrict)
   RSSunrestrict /N- Kunrestrict

Where J = no. extra variables = 2 (Yhat2 and Yhat3)
And kunrestrict = no. of right hand side variables including the constant in the
  regression that includes the higher order powers of the residuals = 8

= 3327.6 - 3315.4 /2 ~ F(2, 12098-8)
   3315.4 /12098 - 8
= 22.2

From F tables, critical value at 5% level F(2, 12090) = F(2, ∞ ) = 3.00

So estimated F > Fcritical
So reject null that model has no omitted variables
Check that automatic version of test gets similar answer.

```
  . ovtest /* this is stat command for the reset test */
```

Ramsey RESET test using powers of the fitted values of lhpay

Ho: model has no omitted variables

\[ F(3, 12089) = 25.20 \]
\[ Prob > F = 0.0000 \]

Yes it does. Note that there are 3 polynomial values of yhat included in stata’s test (look at the degrees of freedom in the F test output above)

2. assuming data read in then 1st regression gives

```
  . reg housepri Incin if year==2
```

```
Source |       SS       df       MS                  Number of obs =     142
---------+------------------------------               F(  1,   140) =   27.73
Model |  4.5871e+10     1  4.5871e+10               Prob > F      =  0.0000
Residual |  2.3159e+11   140  1.6542e+09               R-squared     =  0.1653
---------+------------------------------               Adj R-squared =  0.1594
Total |  2.7746e+11   141  1.9678e+09               Root MSE      =   40672

------------------------------------------------------------------------------
housepri |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Incinera |  -39956.13   7587.677     -5.266   0.000      -54957.38   -24954.89
  _cons |   131902.4    4027.12     32.754   0.000       123940.5    139864.2
------------------------------------------------------------------------------

so might conclude that coefficient on dummy variable indicates that average (mean) house price in area with a waste incinerator was around £40,000 less than in the areas without an incinerator and that this is a statistically significant effect.

But this does not imply that the incinerator is the cause of lower house prices - it may be that house prices were always lower in the area before the incinerator was built.

```
  . reg housepri Incin if year==1
```

```
Source |       SS       df       MS                  Number of obs =     179
---------+------------------------------               F(  1,   177) =   15.74
Model |  1.3636e+10     1  1.3636e+10               Prob > F      =  0.0001
Residual |  1.5332e+11   177   866239953               R-squared     =  0.0817
---------+------------------------------               Adj R-squared =  0.0765
Total |  1.6696e+11   178   937979126               Root MSE      =   29432

------------------------------------------------------------------------------
housepri |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Incinera |  -18824.37   4744.594     -3.968   0.000      -28187.62   -9461.118
  _cons |   82517.23    2653.79     31.094   0.000       77280.09    87754.37
------------------------------------------------------------------------------

The regression for the year before the incinerator was built seems to confirm this. House prices were around £18,800 lower in the years before.

So how to obtain the effect of the incinerator being built? The idea is to compare prices in both areas before and after the construction and examine the
change in the house price gap between the 2 areas. Given a regression of the form:

$$\text{Houseprice} = b_0 + b_1 \text{Year2} + b_2 \text{Incinerator} + b_3 \text{Year2} \times \text{Incinerator} + e$$

It follows that \( \frac{d\text{Houseprice}}{d\text{Incin}} = b_2 \) if year ==1
\[= b_2 + b_3 \text{ if year==2} \]

so \( b_3 \) is the additional effect from year 2, the coefficient of interest

Note also that
the coefficient on the year2 dummy measures the average price across all areas in year 2 relative to year 1
the coefficient on the incinerator dummy measures the average price in the incinerator area relative to other areas averaged across both time periods
the coefficient on the incinerator and year2 interaction dummy measures the change in the average price differential between the two areas in the period after the incinerator was built.

This is the “true” effect of the Incinerator on house prices (the so-called difference-in-difference estimator)

\[ g \text{ inciny2=Incin\times year2} \]

\[ . \text{ reg housepri Incin year2 inciny2} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F(  3,   317) = 58.46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2.1294e+11</td>
<td>3</td>
<td>7.0980e+10</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>3.8491e+11</td>
<td>317</td>
<td>1.2142e+09</td>
<td>Adj R-squared = 0.3562</td>
</tr>
<tr>
<td>Total</td>
<td>5.9785e+11</td>
<td>320</td>
<td>1.8683e+09</td>
<td>Root MSE = 34846</td>
</tr>
</tbody>
</table>

| Housepri | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------|---------|-----------|-------|------|----------------------|
| Inciner y | -18824.37 | 5617.353 | -3.351 | 0.001 | -29876.38 -7772.365  |
| year2    | 49385.15 | 4666.491 | 10.583 | 0.000 | 40203.95 58566.36   |
| inciny2  | -21131.76 | 8591.559 | -2.460 | 0.014 | -38035.44 -4228.08  |
| _cons    | 82517.23  | 3141.95  | 26.263 | 0.000 | 76335.52 88698.94   |

so it would seem that the difference in house prices between the areas grew by £21,000 as a result of the incinerator being built.

You can check this coefficient since it can also be calculated as

\[ (\text{Housepri}_{\text{year2 No Incin}} - \text{Housepri}_{\text{year2 Incin}}) - (\text{Housepri}_{\text{year1 No Incin}} - \text{Housepri}_{\text{year2 Incin}}) \]

\( ie \) the coefficient picks up the change in the differential between the areas.

Check this using Stata command

\[ . \text{ tab year Incin, su(house)} \]

Means, Standard Deviations and Frequencies of housepri

<table>
<thead>
<tr>
<th>year</th>
<th>Inciner</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
</tbody>
</table>
D-in-D = (131902 - 91946) - (82517 - 63692)
= 39956 - 18825
(which are the coefficients on the incinerator dummy in the year specific regressions)
= 21131
(which is the coefficient in the pooled regression)
(Of course other factors might be responsible for this. You might want to include other controls in the regression which could account for the change).

3. Given of a sample of employees, you have data on the number of years of work experience, (YEARS), its square, (YEARS2), and a dummy variable to indicate whether the employee was female or not (FEMALE), together with information on the log of hourly pay measured in pounds, (LHWAGE).

You estimate the following regressions:

(1) \(^\hat{\text{LHWAGE}} = 5.00 + 0.05\text{YEARS} - 0.001\text{YEARS2} - 0.30\text{FEMALE}\)

(2.00) (0.01) (0.002) (0.10)

TSS=10000 ESS=4000 N=124

(2) \(^\hat{\text{LHWAGE}} = 6.00 - 0.25\text{FEMALE}\)

(3.0) (0.10)

TSS=10000 ESS=3000 N=124

where the numbers in brackets are estimated standard errors

i) Interpret the estimated coefficients.

Coeffs. Are semi-elasticities ie give percentage change in pay wrt unit change in rhs variable
Female coeff. says being female means, on average, wages are 30% lower.
Careful with quadratic. Difficult to interpret levels effect independently of squared term, since

\[ d\ln\text{Wage}/d\text{Years} = 0.05-2(0.001)\text{Years} \] (1)
ie effect is not constant

ii) At what level of work experience is pay maximised?

Set (1) = 0 (F.o.C. for maximum)
0=0.05-0.002Years
years=0.05/0.002
iii) Test the hypothesis that years of experience (and its square) have no explanatory power in the model at the 95% level.

Requires F test for sub-set of variables

\[ F = \frac{\text{RSS}_{\text{restrict}} - \text{RSS}_{\text{unrestrict}}}{J} \sim F(J, N - K_{\text{unrestrict}}) \]

Use TSS=ESS+RSS so RSS=TSS-ESS

\[ \frac{10000-4000}{2} = 6000 \text{ in (1) (unrestricted)} \]
\[ \frac{10000-3000}{2} = 7000 \text{ in (2) (restricted)} \]

\[ J = 2 \text{ (2 variables omitted - Years, Years2)} \]
\[ K_{\text{unrestrict}} = 4 \text{ (constant, Years, Years2, female)} \]

so \[ F = \frac{7000 - 6000/2}{6000/124-4} \]
\[ = 500/50 = 10 \]

From F tables, critical value at 5% level \( F(2, 124) = F(2, \infty) = 3.06 \)
So estimated \( F > F_{\text{critical}} \), so reject null that years and years2 have no explanatory power.

4. Given the following information

\[ Cons = 500 + 0.9\text{Income} + 0.3\text{Assets} \text{ for the period 1940-2003} \quad \text{RSS}=700 \]

\[ Cons = 400 + 0.8\text{Income} + 0.2\text{Assets} \text{ for the period 1940-1979} \quad \text{RSS}=350 \]

\[ Cons = 600 + 0.95\text{Income} + 0.35\text{Assets} \text{ for the period 1980-2003} \quad \text{RSS}=250 \]

Test the hypothesis that the coefficients are the same across the 2 sub-periods

Use Chow Test for sample split

\[ F = \frac{\text{RSS}_{\text{restrict}} - \text{RSS}_{\text{unrestrict}}}{J} \sim F(J, N - K_{\text{unrestrict}}) \]

Where \( \text{RSS}_{\text{unrestrict}} = \text{RSS}_{\text{period1}} + \text{RSS}_{\text{period2}} \)

And \( J = \text{number of restricted coefficients} = 3 \text{ (constant, income, assets)} \)

So \[ F = \frac{700 - (350+250)/3}{(350+250)/64-2(3)} \sim F(3, 64-2(3)) \]

(remember with this form of the test there are twice as many coefficients in the unrestricted regressions (income, assets and the constant for the period 1940-79, and a different estimate for income, assets and the constant for the period 1980-03),

so the unrestricted degrees of freedom are
\[ N = N_{40-79} + N_{80-03} = 40 + 24 = 64 \]

and \( k = 2 \times 3 = 6 \)

so \( F = 5 \sim F(3, 58) \)

From table \( F \) critical at 5\% level is 2.76. So estimated \( F > F_{\text{critical}} \). Therefore reject null that coefficients are the same in both time periods. Hence mpc is not constant over time.

5. This is a difference in difference estimation where the “treatment” variable in this case is being female. The key is to be able to interpret each of the coefficients using

Let general model be written as

\[
\ln W = a + a_2 \text{Year}_2 + b_1 \text{Treatment Dummy} + b_2 \text{Year}_2 \times \text{Treatment Dummy}
\]

If \( \text{Year}_2 = 0 \) and \( \text{Treatment Dummy} = 0 \), \( \ln W = a \)
If \( \text{Year}_2 = 0 \) and \( \text{Treatment Dummy} = 1 \), \( \ln W = a + b_1 \)
If \( \text{Year}_2 = 1 \) and \( \text{Treatment Dummy} = 0 \), \( \ln W = a + a_2 \)
If \( \text{Year}_2 = 0 \) and \( \text{Treatment Dummy} = 1 \), \( \ln W = a + a_2 + b_1 + b_2 \)

Compare with

\[
\ln (\text{WAGE}) = 2.00 + 0.05 \text{YEAR}_2 - 0.250 \text{FEMALE} + 0.03 \text{FEMALE} \times \text{YEAR}_2
\]

(1.00) (0.01) (0.002) (0.02)

so coefficient \( a = 2.00 \) is effect of “control” in base period = 2.00 log points (around £7.38 an hour, since \( \exp(2) = 7.38 \))

coefficient \( a_2 = \) change in wages of control group in 2nd period relative to 1st = 0.05
Since this is a semi-log model, this means wages rose by 5\% in the 2nd period compared to the 1st

coefficient \( b_1 = -0.250 \) is effect of “treatment” in base period = so women earned around around 25\% an hour less in the base period, since this is a semi-log model

coefficient \( b_2 = \) change in wages of treatment group relative to that of control group between 1st and in 2nd period = 0.03
Since this is a semi-log model, this means wages for treatment group (women) rose by 3\% more than that or control group (men) in the 2nd period compared to the 1st
But estimate is not statistically significantly different from zero (t = 0.03/0.02 = 1.5)

So policy had no discernable effect

6. In April 2000 the UK government introduced the Working Families Tax Credit aimed at increasing the income in work relative to out of work for groups of traditionally low paid individuals with children. In addition financial help was also given toward child care.
If successful the scheme could have been expected to increase the hours worked of those who benefited most from the scheme—namely single parents. By comparing hours of worked for this group before and after the change with a suitable control group, it should be possible to obtain a difference in difference estimate of the policy effect.

The following example uses other single childless women as a control group.
. tab year, g(y)
/* set up year dummies. Stata will create two dummy variables
y1=1 if year=1998, = 0 otherwise
y2=1 if year=2000, = 0 otherwise */
.
.g lonepy2=lonep*y2
/* create interaction variable */
.
.reg hours lonep if year==1998
Source |       SS       df       MS              Number of obs =    3188
-------------+------------------------------           F(  1,  3186) =  857.65
Model |  235927.871     1  235927.871           Prob > F      =  0.0000
Residual |  876422.651  3186  275.085578           R-squared     =  0.2121
-------------+------------------------------           Adj R-squared =  0.2119
Total |  1112350.52  3187  349.027462           Root MSE      =  16.586
------------------------------------------------------------------------------
hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lonep |  -17.2959  .5905917   -29.29   0.000    -18.45388   -16.13792
       _cons |   28.7082   .3956836    72.55   0.000     27.93238    29.48402
------------------------------------------------------------------------------
.
.reg hours lonep if year==2000
Source |       SS       df       MS              Number of obs =    3124
-------------+------------------------------           F(  1,  3122) =  756.20
Model |  189616.677     1  189616.677           Prob > F      =  0.0000
Residual |   782842.48  3122  250.750314           R-squared     =  0.1950
-------------+------------------------------           Adj R-squared =  0.1947
Total |  972459.157  3123  311.386218           Root MSE      =  15.835
------------------------------------------------------------------------------
hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lonep |  -15.63756  .5686585   -27.50   0.000    -16.75254   -14.52258
       _cons |   27.4026   .3847371    71.22   0.000     26.64823    28.15696
------------------------------------------------------------------------------

The coefficient on lone parents gives the difference in average hours worked between lone parents and the control group for the relevant year.

Comparing the lone parent coefficient across periods, lone parents worked 17.3 hours less than other single women in 1998 before the policy, (28.7-17.3 = 11.4 hours a week for single parents on average in 1998) and 15.6 hours less than other single women immediately after the introduction of WFTC, (27.4-15.6 = 11.8 hours a week for lone parents in 2000, on average).

So the change (difference in difference)
= -15.6 – (-17.3) = 1.7
= (Hours\textsuperscript{LonePar}_{2000} - Hours\textsuperscript{LonePar}_{1998}) - (Hours\textsuperscript{Single}_{2000} – Hours\textsuperscript{Single}_{1998})
= (11.8-11.4) - (27.4 – 28.7) = 0.4 – (-1.3) = 1.7

which suggests lone parents worked relatively about (net) 1.7 hours more as a result of the policy. (Note that hours worked actually fell for the control group).

To obtain standard errors, pool the data and estimate the following
6. The change is also statistically significant

To analyse effect of congestion charge need to compare change in traffic volumes for group affected by congestion charge with change for those not (not sufficient to do single period regression like

\[ \text{Volume} = a + b_1 \text{London driver} \]

Since treatment group may be systematically different (eg London commuters typically drive longer distances than other commuters. Need to net out this difference otherwise would attribute this effect (wrongly) to congestion charge)

Hence need to compare change in coefficients at two points in time before & after the introduction of the charge

\[ \text{Volume}_1 = a + b_{11} \text{London driver} \quad \text{in period 1} \]
\[ \text{Volume}_2 = a_2 + b_{12} \text{London driver} \quad \text{in period 2} \]

Or its equivalent

\[ \text{Volume} = a + a_2 \text{Year}_2 + b_1 \text{London driver} + b_2 \text{Year}_2 \times \text{London driver} \]